STAT 722/422: Predictive Analytics for Business
Syllabus, Fall 2016

Course Info

- **Course times**:
  Section 402: MW 9–10:30, 10/24–12/12, JMHH 255
  Section 404: MW 12–1:30, 10/24–12/12, JMHH F70
- **Instructor**: Yuancheng Zhu, `zhuyuanc@wharton.upenn.edu`
  Office: JMHH 455
  Office hours: MW 3–4
- **Course assistant**: Colman Humphrey, `chump@wharton.upenn.edu`
  Office hours: T 2–3, JMHH F92
  Th 2–3, JMHH F86
- **Pre-requisite**: STAT 613/621

Course Overview

This seven-week course introduces students to the statistical techniques that extend the ideas of regression analysis introduced in STAT 613. Digressing from traditional approaches that focus on carefully modeling how one or two chosen measurements relate to a response, we will take a "modern" approach applicable to managerial decision making in the presence of large data sets.

After a brief review of least squares regression, we will round out our regression toolbox by learning how to build models for predicting categorical responses. Equipped with a solid foundation, we will switch our approach to the point of view of predictive modeling using automatic tools. The name of the game in predictive modeling is to be able to predict the behavior of new data. If, for example, we can show a bank how to predict who will default on a loan better than their existing system, the bank can increase profits. Similarly, if we help a company identify those in the market most interested in its products, then it can construct a much more focused product launch.

While our focus will center around prediction, managers still want to understand where their forecasts come from as there is still much hesitance in trusting pure "black-boxes" in many business scenarios. Hence, we will constantly explore and emphasize the trade-off between prediction power and model interpretability.

As the business world rapidly progresses towards a paradigm of data-driven decision making, the primary goal of this course is on understanding both the power and limitations of regression analysis. The course is designed to allow future managers–both data scientists and not–to communicate effectively with the data science team within an organization.

We are going to let software do the number crunching for us–our value-add comes from how to choose to tackle the problem and what insights we can draw from the model results.

Grading

- **Homework**: 60%

  There will be 5 weekly assignments.

  All assignments except the 5th one are due Wednesday of that week. You can submit your homework either in class or to JMHH 455 during the office hour (3–4pm) on Wednesday.

  The 5th homework will contain two prediction problem and no written solution is required. The score will depend solely on the prediction submitted electronically. More details will be provided when the homework is out.

  Late policy: Late assignments will be penalized at 10% of the maximum grade per day for up to two days and are ineligible to be handed in for credit after this time. Please place late homework directly in my mailbox in the Statistics Department.

  Collaboration policy: Working together on homework is allowed and encouraged. However, students must write up their homework solutions by themselves. Names of collaborating students should be provided on the front page of each homework write-up.

- **Final exam**: 40%

  An in-class, open-book and open-computer final exam will be held in the last lecture.

  The exam will consist of both conceptual questions and small data analysis tasks. Datasets to be used will be posted one day prior to the exam. Students are supposed to bring their own laptops and perform data analysis using software during the exam.

Course Materials

- **Lecture slides**

  The lecture slides will be the primary learning guide for the course and should be fairly complete. The slides will be posted by the morning of lecture (sometimes the night before) and you are encouraged to bring them to class as an aid.

- **Textbook**

  There is no required textbook for this course. Below are some recommended textbooks that cover the material presented in the context of much more extensive treatments of advanced modeling.

  - James, Witten, Hastie and Tibshirani. *An Introduction to Statistical Learning with Applications in R.* (A free copy can be obtained at `http://www-bcf.usc.edu/~gareth/ISL/`)
  - Kuhn an Johnson. *Applied Predictive Modeling.*

- **Computer software**

  The software to be used in the course is **JMP 12 Pro**.

  JMP 12 Pro can be downloaded from Canvas. Mac and Windows versions are available. Instructions are provided on Canvas. The part of manuals for JMP that are mostly related to

the course can be found at `http://www.jmp.com/support/help/Fitting_Linear_Models.shtml`.

JMP session: Tuesday, Oct 25, 2–3 in JMHH F92
Thursday, Oct 27, 2–3 in JMHH F86

The primary motivation for using JMP in class is to ensure that lack of prior programming experience is not prohibitive to any student in the class. That being said, you may use any and all software packages you like to complete the assignments in the class, as would likely be the case in many real-world data analysis scenarios.

Schedule (very tentative)

Date	Day	Topics	Assignment
Oct 24	Mon	Introduction; multiple linear regression	
Oct 26	Wed	Multiple linear regression; logistic regression	Assn 1 out
Oct 31	Mon	Logistic regression cont.	
Nov 2	Wed	Classification & k nearest neighbors	Assn 1 due, Assn 2 out
Nov 7	Mon	Out-of-sample validation	
Nov 9	Wed	k-fold cross-validation	Assn 2 due, Assn 3 out
Nov 14	Mon	Subset selection	
Nov 16	Wed	Shrinkage methods	Assn 3 due, Assn 4 out
Nov 21	Mon	Model selection cont.	
Nov 28	Mon	Case studies	Assn 5 out
Nov 30	Wed	Decision tree	Assn 4 due
Dec 5	Mon	Random forest	
Dec 7	Wed	Review	Assn 5 due
Dec 12	Mon	Final exam	